DOCUMENT RESUME

ED 335 359 TM 016 177

AUTHOR Steinberg, Wendy J.
TITLE Differences between Novice and Expert Knowledge Structure, Pre- and Post-Training, in a Statistics and Test Theory Domain.
PUB DATE Nov 90
NOTE 27p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, November 1, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *Cognitive Structures; *Comparative Testing; *Government Employees; Job Training; *Knowledge Level; Mathematics Instruction; Multidimensional Scaling; Multiple Choice Tests; Occupational Tests; Pretests Posttests; Professional Education; Rating Scales; State Government; *Statistics; Test Items; *Test Theory
IDENTIFIERS Experts; New York

ABSTRACT

The purpose of this study was to examine the nature and degree of differences in expert versus novice knowledge structures, both before and after training, when judging the similarity of multiple-choice test items within a statistics and test theory (STT) domain. Subjects were employees of the Testing Division of the New York State Department of Civil Service (NYDCS), of whom 8 were trainees. Approximately 2 weeks prior to taking the agency training course in elementary STT, a 45-item multiple-choice domain-specific pretest was administered to the 8 trainees, as well as to 10 experienced NYDCS staff members. Scores on the test were then used as the basis for identifying novices and experts. The final subject pool (6 novices and 6 experts) was then administered a subset of 18 items from the pretest covering descriptive statistics; inferential statistics and experimental design; and validity, reliability, and test theory. Subjects judged on a 7-point scale how similar they perceived 153 different item pairs to be. At the end of the course, novices again sorted the 153 pairs of items using the 7-point scale. Subjects' multidimensional scaling judgments were entered into matrices and analyzed via INDSCAL, and an analysis of angular variation was computed. Prior to training, no significant difference between the two subject groups was found, although at that point novices shared only one of the experts' three dimensions, while, after training the groups shared two dimensions. The finding of no significant difference between the groups prior to training conflicts with logic, and the paper concludes with a discussion of possible reasons for the finding of non-significance. Nine transparencies illustrating the study design and results are appended. (RLC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

WENDY J. STEINBERG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 325356

DIFFERENCES BETWEEN NOVICE AND EXPERT KNOWLEDGE STRUCTURE, PRE AND POST-TRAINING, IN A STATISTICS AND TEST THEORY DOMAIN

Wendy J. Steinberg

State University of New York at Albany

Paper presented at the annual meeting of the
Northeastern Educational Research Association
November 1, 1990, Ellenville, NY

BEST COPY AVAILABLE

TM 016 177

DIFFERENCES BETWEEN NOVICE AND EXPERT KNOWLEDGE STRUCTURE, PRE AND POST-TRAINING, IN A STATISTICS AND TEST THEORY DOMAIN

Statement of Problem

The purpose of this study was to examine the nature and degree of differences in expert versus novice knowledge structures, both pre and post-training, when judging the similarity of multiple-choice test items within a statistics and test theory domain. Affirmative answers were expected to the following specific questions:

1. Prior to training, do novices and experts use different dimensions/concepts?

2. Prior to training, do novices and experts give different weights (saliency) to their shared dimensions/concepts?

3. If yes, is the difference statistically significant?

4. Finally, what are the answers to the above questions when assessed after training rather than before? In other words,

do novices' mental representations change in the direction of experts' as a result of training, and if so, is the amount of change significant?

Conceptual Framework

Multidimensional Scaling (MDS) is a technique for capturing and representing the "structure" of a person's (or group of persons') perceptual or semantic knowledge or affective preferences, within a particular content domain. The data base consists of subjects' similarity or dissimilarity judgments for all possible pairs of a set of items or objects within the domain of interest. For example, a market researcher might ask subjects to judge the similarity of all possible pairs of soft drinks. The educational psychologist, on the other hand, may ask subjects to judge all possible pairs of a set of concepts within the subject area being taught. The dimensions, traits, or features upon which the ratings are to be made are not, however, specified a priori to the subjects. Instead, by looking post hoc at the items or objects which subjects group together as most similar vs. most dissimilar, the investigator can, much as in factor analysis, intuitively "label" the dimension, trait, or feature that the items or objects have in common. The intent, however, is not to label the items themselves, but rather to describe the underlying latent concepts, ideas, and perceptions that the

persons doing the judging are employing. The logic is that the relationships which the subjects impose on the items or objects is a function of, and therefore reflects, the relational connectedness of such items or objects in those persons' minds. Hence, the intent is to "peek into the subjects' minds", so to speak, to see how the information stored therein is conceptually organized.

Multidimensional scaling, while primarily descriptive, does have the potential to constrain, and thereby inform, theorizing about both process and behavior. For example, a direct relationship has been found between the reaction time for subjects to decide whether or not the objects in a pair are from the same conceptual class, and the Euclidean distance of those objects from their concepts' centroids in multidimensional space, and that finding was then used to propose a sequential mode of information processing (Rips, Shoben, and Smith, 1973). Others have noted that subjects' choice of the element to complete an analogy is a function of the element's euclidean distance from the correct answer's ideal point (centroid) in multidimensional space, and that the mathematics in Luce's choice rule is able to predict the distribution of subjects' choices with a high degree of accuracy (Rumelhart and Abrahamsen, 1973). MDS can also be used to elucidate the mental processes that underlie scores on traditional psychometric tests. For example, an MDS assessment of traditionally tested examinees may reveal that for some examinees the test is primarily a verbal task, while for other

examinees the same test items are primarily a visual-spatial task. On a group basis, such findings could provide insight into test bias, as well as suggest instructional (treatment) modifications.

Probably the most popular use to which MDS has been put within the field of education, however, is to describe the nature of the differences in concept/knowledge organization between two groups--say, novices versus experts--and then to determine the degree and nature of the change in the novices' representation toward the experts' representation as the result of various types or amounts of instructional intervention. The usual findings are that the novices' representation is more scattered than the experts' and reflects non-essential "surface" features of the objects being studied: for example, the presence of levers or pulleys in physics problems. Experts, on the other hand, have a representation which is more spatially compact and which reflects principles-based concepts: for example, the presence of velocity or resistance features in physics problems (Chi, Feltovich, and Glaser, 1981). The more interesting finding is that the novices' representation converges on the experts' representation as a function of instruction, and moreso in principles-based instruction.

While these types of studies provide interesting descriptions of the differences between novices and experts, what is missing is some type of external validation of the accuracy of

the labelled dimensions, as well as a statistical test of the significance of the difference between the two groups. Procedures for conducting such validations and for significance testing do exist and are well-described in the text by Schiffman, Reynolds, and Young (1981). Unfortunately, few studies in the literature have employed them. Thus, the reader is left wondering about the accuracy of the researcher's dimension labels as well as the statistical significance of the observed novice-expert differences. In addition, the set of stimuli items to be rated have in many cases consisted of single elements--for example, colors, fruits, or soft drinks--making it difficult to see the connection to the complex problem-solving processes and knowledge structures involved in education and learning.

The current study examines the differences between novices and experts in a particular content domain--elementary statistics and test theory--by use of an externally validated description of the dimensions and a statistical test of the amount of observed difference between the groups. In addition, the stimuli consist of multiple choice items, i.e., problems, rather than single words. Hence, the educational applications are immediately evident.

Methodology

The subject pool for this experiment consisted of

employees of the Testing Division of the New York State Department of Civil Service, the State's centralized personnel testing and selection agency. Trainees enter the agency with no previous coursework or experience in either statistics or test theory, but rather, receive intensive coursework and on-the-job training over a two-year period. One aspect of that training is a 10-week (30-hour) course in elementary statistics and test theory. At the time of this study, there were 8 Trainees in the agency.

Approximately 2 weeks prior to course intruction, a multiple choice (M/C) domain-specific "pre-test" was administered to the 8 Trainees in the course, as well as to 10 experienced staff. The experienced staff were deliberately chosen to relect a range of medium to high statistical competence, based upon the author's personal knowledge of their on-the-job abilities. (That the author's judgment was accurate was reflected by a .99 Pearson correlation between ranked multiple-choice test scores and external criterion ratings of statistical/test theory competence, as well as the location of 6 out of the 8 Trainees in the bottom 8 positions of the scores distribution.) The M/C test consisted of 45 items, broken down into 15 items each in the areas of: (a) descriptive statistics, (b) inferential statistics and experimental design, and (c) validity, reliability, and test theory.

Scores on the M/C test were then used as the basis for

deciding which subjects were ultimately placed in the "novice" versus "expert" groups -- i.e., the group classifications were empirically based. To better differentiate the two groups, the subsequent MDS analysis was limited to the six lowest and six highest M/C scorers. Hence, the final subject pool consisted of 12 subjects.

Immediately following completion of the M/C test, the 12 subjects were each given a subset of the multiple-choice items on which to perform the multidimensional scaling sorting task--6 items each from the 3 content areas, for a total of 18 items. All subjects received the same 18 items. There were thus [(18)(17)]/2, or a total of 153 different pairs of items which the subjects had to judge. For each pair of items, subjects were required to judge how "similar" they perceived the items to be. Similarity judgements were to be placed on a 7-point scale, with "1" being very similar and "7" being very different. At the completion of the training course, subjects again sorted the same 153 pairs of items, using the same 1-7 scale.

Subjects' MDS judgments were entered into matrices, one for each subject. The data was then input and analysed using the SPSS version of INDSCAL and, where necessary, manual calculation. Analysis of Angular Variation (ANAVA) was computed through use of a personally written Matrix Algebra Tool System (MATS) program.

Results

Stress indices (badness of fit) for 4, 3, and 2 dimensional solutions in the pre-training condition were .168, .209, and .306 for novices, and .155, .198, and .277 for experts. R-squared values (accountable variance) for the same solutions were .599, .528, and .475 for novices, and .697, .676, and .632 for experts. The 4-dimensional solutions were not intuitively interpretable. Therefore, analysis was limited to the 3-dimensional solutions.

Prior to training, the dimensions used by novices versus experts did, indeed, appear to differ. Looking at the content, format, and underlying principles of the items that clustered together, novices' mental representations appeared to be organized by: (a) Inferring the effect on a statistic of a change in conditions vs. Selecting and computing a statistic, (b) Descriptive and inferential statistics vs. Correlation and test theory,(c) an uninterpretable dimension. Experts, on the other hand, used the following dimensions: (a) Experimental Design vs. Not experimental design, (b) Descriptive and inferential statistics vs. Correlation and test theory, (c) Factual recall vs. Judgment and application.

On a post-training basis, stress values for 4, 3, and 2 dimensions were .148, .200, and .281 for novices. R-squared values were .688, .591, and .528. Experts were not reassessed,

since they did not receive the training. (The sorting task is so time-consuming that the threat to internal validity due to this omission was not judged to be high enough to justify again imposing the task on the experts.) The novice dimensions for the 3-dimensional solution were: (a) Selecting & computing a statistic vs. Inferring the effect on a statistic of a change in conditions, (b) Descriptive and inferential statistics vs. Correlation and test theory, (c) Factual recall vs. Judgment and application. Again, experts were not reassessed, since they did not receive the training. Note that the novices' representations did become more "expert-like" after training than they had been prior to such training: Prior to training, novices shared only one of the experts' three dimensions, but after training they shared two.

The above analyses were performed on novices and experts separately. However, to examine differences between novices versus experts on a statistical rather than merely intuitive basis, all subjects--novices and experts combined--must be loaded into the same dimensional "space". This is because one cannot compare the statistical significance of dimension saliance (weight) for different groups unless the dimensions that the two groups used were the same. When all subjects were combined, the new stress values across 4, 3, and 2 dimensional solutions were .170, .221, and .305. The R-squared values were .568, .528, and .481. Again, for reasons of interpretability, analysis was limited to the 3-dimensional solution. The dimensions/concepts

were best labelled as:

1. Descriptive & inferential statistics vs. Correlation and test theory

2. Computing or selecting a statistic vs. Understanding the effects on a statistic of a change in conditions

3. Numerically formatted vs. Not numerically formatted

Note that these dimensions were still only intuitively labelled. To validate the accuracy of the dimension labels, the 18 items were given to an external group of experts (i.e., high-ranking agency psychometric staff who were not a part of the original MDS task), along with the dimension labels. These experts were then asked to rate the degree to which the items actually "contained" or reflected those dimensions. The correlation between the items' "actual" versus "labeled" dimensions for the first dimension was .836; for the second, .317; and for the third, .498. Although the latter two values appear to be somewhat low, it should be noted that the task before the external group of experts was to assess the r- ationship between items and dimensional labels separately for the three dimensions, while the task before the original subjects was to sort items according to whatever dimensions they were simultaneously imposing. This difference in task constrains the possible correlation. Thus, the correlations should be viewed as

lower bounds: the true correlations are "at least" as great as the observed correlations. Hence, there is reasonable support for each dimension label.

Finally, given a 3-dimensional solution common to both novices and experts, the statistical significance of the difference in dimension saliency between the two groups was assessed. Because the technique for doing this is not widely known, a short digression to describe the technique is appropriate. In multidimensional scaling, the "subject space" indicates the weight, or saliency, of the various dimensions for the individual. However, because subjects' dimension weights are expressed in vectors rather than points, and both the dimension axis and the subject's weight vector share the same origin, the dimension saliency for that individual is indicated not by distance from the dimension axis but by angle from the dimension axis. Therefore, linearly-based statistics do not apply. Instead, one must use an obscure branch of statistics known as "directional" statistics (see Mardia, 1972). The technique for testing the significance of differences between groups when the data is measured in terms of variation in vector angles is called "analysis of angular variation", or ANAVA for short. Its use permits one to investigate how individuals or groups differ in the way they use the dimensions within the grouped stimulus space. It is distributed and interpreted as the familiar f. In the current study, the degrees of freedom are 2 and 20, because of the multiplicative effect of dimensionality. The observed f

(2,20) = 1.68, p>.05. Hence, there was no significant difference between novices and experts prior to training.

Because there was no significant difference between the groups prior to training, and the training should, if anything, cause the novices to become more like the experts, (as, indeed, the separate dimension labels described above indicated it had), the investigation into whether or not there was a significant difference between the groups after training was not conducted.

Discussion

The finding of no significant difference between groups prior to training conflicts with logic, since the novices were nearly all completely naive statistically, and the experts each competent. Indeed, the multiple choice test which was used as an objective criterion for sorting the original pool of subjects into the two groups bore that out: Each of the six subjects classified as a novice on the basis of low score on the test of statistical knowledge was in fact a new trainee to the agency, and each of the six subjects classified as an expert on the basis of high test score was in fact an experienced and psychometrically respected staff member. Clearly, there was a difference between the groups at least in factual knowledge and its applications, if not in the structure of that knowledge.

One explanation for non-significance might be that the domain is one which is not dimensionally structured in the first place. If the model does not fit the domain, it will misrepresent the degree of relationship between novice-expert structures, much as a Pearson correlation will misrepresent the degree of relationship when data is curvilinear. Research has shown, for example, that dimensional models are often better descriptions of knowledge structure when the domain is perceptual or sensory than when it is semantic; Semantic domains are often better described through hierarchical or tree models (Tversky and Hutchinson, 1986). To determine if a dimensional model was a poor choice for the domain, "centrality" and "reciprocity" statistics, which are diagnostic aids in choosing between the types of models just mentioned, were computed. Centrality was 3.23 and reciprocity was 3.85. While those values are beyond the purely theoretical upper bounds for dimensional models, they were within the range which Tversky and Hutchinson's work indicates are typical of real data fitting dimensional models. Hence, the choice of a dimensional model was not unreasonable.

A second explanation for non-significance might lie in the small sample size. However, it is unlikely that sample size alone explains the non-significance. Because of the time-consuming nature of the sorting task for the subjects (object pairs=[(k)(k-1)]/2) and the need to data-enter separate matrices of judgments for each subject for the sorted items, most MDS research consists of relatively few subjects and relatively

few items: 8 to 20 subjects and 12 to 20 items. Both the number of subjects and the number of items in this study were typical of the MDS literature.

A third explanation for non-significance might lie with the stimuli language: Several items contained the words "correlation", "reliability", etc., and those surface words unfortunately sometimes inadvertently described the items' deeper underlying principles. Therefore, the words alone may have prompted the novices to sort the items on the same dimensions as the experts, even if only the experts were sorting on the basis of a deeper understanding of the items' underlying principles. However, such flaws existed in only a few of items; it is unknown whether the "noise" created by those several flawed items was sufficient to mask a real difference between groups.

A final possible explanation may lie in the complexity of the stimulus items. As previously mentioned, most MDS research involves single words and simple concepts (e.g., the names of foods or animals). Little research has been done using complex stimuli such as the multiple choice items used in this study. Perhaps either the format of the items or their conceptual complexity interacts with, or even determines, knowledge structure. While this is something educators should want to know, the current study was not designed to address that possibility.

While the results were disappointing, this study is nonetheless of interest because of its application to a different domain than those previously or typically investigated and because of the use of validation and inferential techniques which advance MDS beyond intuitive description. Further studies are suggested to rule out the competing explanations suggested above.

References

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Representation of physics knowledge by experts and novices. Cognitive Science, 5 121-152.

Mardia, K. V., (1972). Statistics of directional data. New York, NY: Academic Press.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and verification of semantic relationship. Journal of Verbal Learning and Verbal Behavior, 12, 1-20.

Rumelhart, D. E., & Abrahamson, A. A., (1973). A model for analogical reasoning. Cognitive Psychology, 5, 1-28.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). Introduction to multidimensional scaling: Theory, methods, and applications. Orlando, FL: Academic Press.

Tversky, A., & Hutchinson, J. W., (1986). Nearest neighbor analysis of psychological spaces. Psychological Review, 93(1), 3-22.

RESEARCH QUESTIONS

1. Prior to training, do novices and experts use different dimensions/concepts?

2. Prior to training, do novices and experts give different weights (saliency) to their shared dimensions/concepts?

3. If yes, is the difference statistically significant?

4. Finally, what are the answers to the above questions when assessed after training rather than before? In other words, do novices' mental representations change in the direction of experts' as a result of training, and if so, is the amount of change significant?

METHODOLOGY

1. 45-Item Multiple-Choice Pre-Test
   (15 items in each of 3 Subtest areas)
   Given to 8 Trainees and 10 Experienced Staff

2. 6 Lowest Scorers and 6 Highest Scorers

   "Novices" "Experts"

3. 18 Multiple-Choice Items Given as a Multidimensional Scaling Sorting Task: [(18)(17)/2] Pairs

4. Resulting Dimensions Labelled Separately for Novices vs. Experts

5. Training Course Given

6. Multidimensional Scaling Sorting Task Repeated and Relabelled

PRE-TRAINING

STRESS

	Novices	Experts
4-dim	.168	.155
3-dim	.209	.198
2-dim	.306	.177

R-SQUARED

	Novices	Experts
4-dim	.599	.697
3-dim	.528	.676
2-dim	.475	.632

POST-TRAINING

STRESS

	Novices
4-dim	.148
3-dim	.200
2-dim	.281

R-SQUARED

	Novices
4-dim	.688
3-dim	.591
2-dim	.528

PRE-TRAINING

Novices

1. Inferring the effect on a statistic of a change in conditions
   vs.
   Selecting and computing a statistic

2. Descriptive and inferential statistics
   vs.
   Correlation and test theory

3. [an uninterpretable dimension]

Experts

1. Experimental design
   vs.
   NOT experimental design

2. Descriptive and inferential statistics
   vs.
   Correlation and test theory

3. Factual recall
   vs.
   Judgment and application

POST-TRAINING

Novices

1. Inferring the effect on a statistic of a change in conditions
   vs.
   Selecting and computing a statistic

2. Descriptive and inferential statistics
   vs.
   Correlation and test theory

3. Factual Recall
   vs.
   Judgment and application

METHODOLOGY

1. Multidimensional Scaling Sorting Task Data Reanalyzed for Novices and Experts Combined; Dimensions Relabelled

2. Dimension Labels Externally Validated

3. Analysis of Angular Variation Computed Between Novices and Experts

4. Dimensional vs. Hierarchical Paradigms Tested

PRE-TRAINING

Novices and Experts combined

STRESS

4-dim	.170
3-dim	.221
2-dim	.305

R-SQUARED

4-dim	.568
3-dim	.528
2-dim	.481

DIMENSIONS (3-dim)

1. Descriptive and inferential statistics
   vs.
   Correlation and test theory

2. Understanding the effects on a statistic of a change in conditions
   vs.
   Computing or selecting a statistic

3. Numerically formatted
   vs.
   NOT numerically formatted

DIMENSION VALIDATION

Dim #1	.836
Dim #2	.317
Dim #3	.498

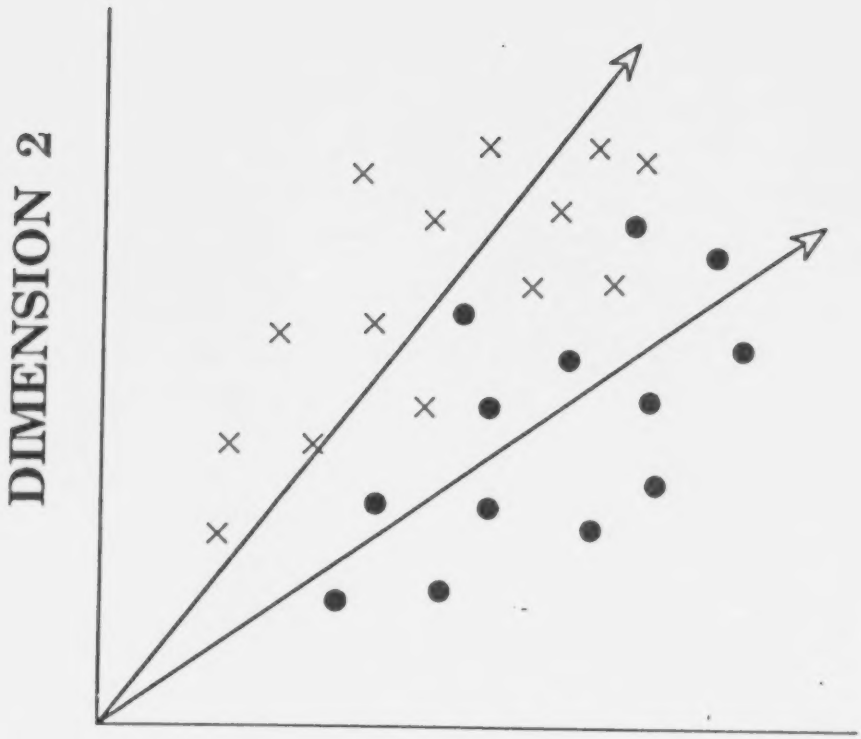
ANAVA

$F(2,20) = 1.68 \quad p > .05$

DIMENSIONAL vs. HIERARCHICAL

Centrality = 3.23

Recriprocity = 3.85



● = novice subject (hypothetical)

× = expert subject (hypothetical)